

# Introduction to Biostatistics

## Exercises on Lecture 1

September 18, 2022

## R Codes: Copy codes to R console window

- Creating a vector: `x<-c(1,2,3,4)`
- Calculating mean & sd: `mean(x)` & `sd(x)`
- Calculating summary measures: `summary(x)`
- Calculating SEM: `sd(x)/sqrt(length(x))`
- Calculating shape measures (need to install & call `moments` package)
  - `install.packages("moments")`
  - `library(moments)`
  - `kurtosis(x)` & `skewness(x)`

## R Codes: Copy codes to R console window

- Generating normally distributed data: `rnorm(n, mean,sd)`
- **Example:** Generating a random sample of size 100 from a normal distribution with mean=0 and sd=1
  - `y<-rnorm(n=100, mean=0,sd=1)`
- Graphical presentation of quantitative data
  - `hist(y)` & `boxplot(y)`
- Graphical parameter **mfrow**—specify the number of subplot
  - `par(mfrow=c(1,2))` set the plotting area into a 1\*2 array

- **Question 1.** Simulate random samples of size 5, 20, 50, 70, 100, 200, 500 & 1000 from a normal distribution with mean=5 & SD=2.
  - For each n, plot a histogram & compute mean, SD, SEM, skewness, kurtosis
  - Discuss the impact of increasing sample size on the shape of the histogram and the descriptive measures
  - What percentage of the values would you expect to lie within 2 standard deviations of the mean? Which rule does a better job in summarizing the data: Empirical rule or Chebyshev's rule? Why?

- **Question 2.** A study was conducted to investigate the physical activity of college students as assessed by number of miles walked per week. Listed below are data from 9 students who participated in the study: 5,6,10,12, 13, 14, 15, 16, 40
  - Find mean, median, mode, range, IQR, SD
  - Compute and interpret the skewness and kurtosis
  - Which measure of central tendency is appropriate: mean or the median or both? Why?
  - Which measure of dispersion would you report: range or interquartile range or both? Why?

- **Question 3.** What is the difference between probability and non-probability sampling techniques? Explain the differences among the three commonly used probability sampling methods? State the conditions under which we use each method?
- **Question 4.** Why do we want to know the SE? What is the difference between SD & SEM? When can SEM be used instead of SD?